

PERIODICAL STACKS

NEWSPAPER FILE

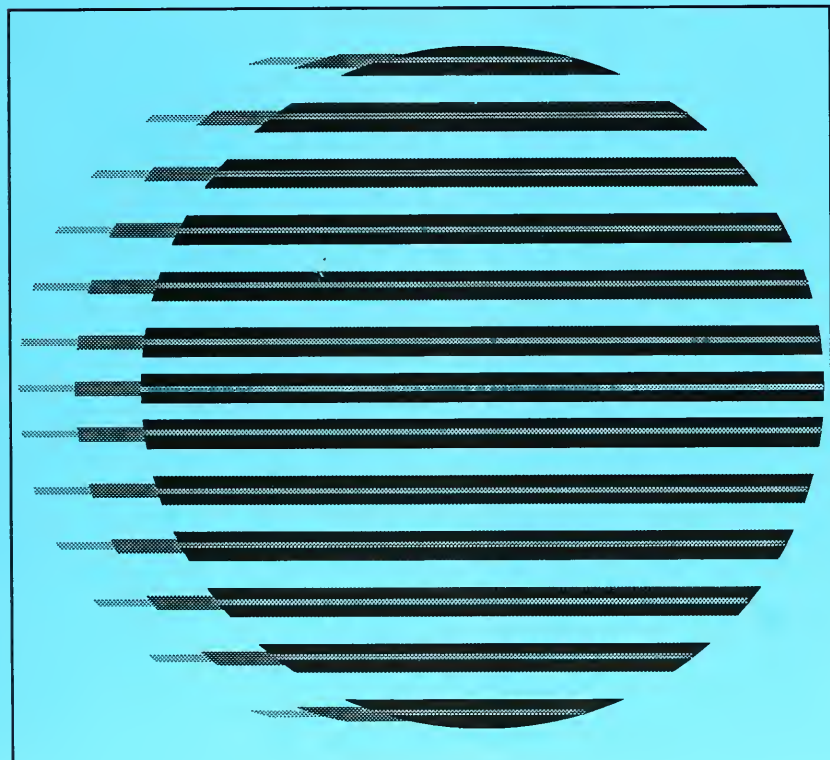
IASSIST

Q U A R T E R L Y

VOLUME 19

Winter 1995

NUMBER 4





Digitized by the Internet Archive
in 2010 with funding from
University of North Carolina at Chapel Hill

<http://www.archive.org/details/iassistquarterly194inte>

IASSIST QUARTERLY



The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines" *Journal of the American Society for Information Science* 30(2):77-82, March 1979. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Laura Bartolo, Libraries & Media Services, Kent State University, Kent, Ohio 44242. (216) 672-3024. Email: LBARTOLO@KENTVM.KENT.EDU. Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCLD

Title: Newsletter - International Association for
Social Science Information Service and
Technology

ISSN - United States: 0739-1137 Copyright 1985 by
IASSIST. All rights reserved.

CONTENTS

Volume 19

Number 4

Winter 1995

FEATURES

- 4** Scientific Data and Social Science Data Libraries
by David Barber & Jan Zauha
- 7** Building an Archive of U.S. Census Related Data Products
by John Blodgett
- 12** Developing a Scottish Migration Monitor: a co-operative approach
by A. McCleery, E. Forster, H. Ewington & P. Burnhill
- 24** Public Access to Large Data Sets in a Depository Library
by Juri Stratford
- 26** Preserving Scientific Information on the Physical Universe
by Kenneth Thibodeau

Scientific Data and Social Science Data Libraries

by David Barber, Coordinator of Information Technology & Jan Zauha, Documents and Data Support Librarian Graduate Library University of Michigan*

There is a vast amount of quantitative information available in electronic form. Social science data makes up less than half that amount. The other, larger half is scientific data. While university libraries have made a considerable investment in social science data, little has been done about scientific data. If administrators, librarians, or others believed that more attention should be paid to scientific data, one of the suggestions that might naturally arise is that social science data specialists should be involved. Though some common ground between the two areas should be acknowledged, the existence of very substantial differences must also be recognized. Those differences are especially significant because coping with them will require an investment of staff and financial resources by the data library.

What Is Scientific Data?

There are many ways to typify scientific data. It could be identified in terms of the disciplines that create and use it. Scientific data is then the data of chemists, geologists, physicists, and others. This type of data could also be identified by the type of phenomena that are measured. Scientific data measures physical, not social phenomena. For example, there are the kind of scientific measurements made in labs or in the field that yield a single number or a series of numbers describing physical phenomena like atomic weights, inches of precipitation, and voltage. There are measurements made of radiation reflected by or passing through objects such as are made by remote sensing or MRI scans. In addition, scientific data may be the result of a series of equations that model the behavior of a turbulent fluid and the numbers that are produced when that simulation is run.

Scientific data information is found in many locations. It is available from research institutes, commercial firms, and governmental data archives. It is often stored on single-user computer systems either on a scientist's desktop or in a science library. The data are also often stored on departmental servers.

Social Science and Science Data: Is There a Connection?

There are things that social science data libraries share with the world of scientific data. The most basic, and perhaps important, similarity in the minds of non-data users is that both are numeric types of information. It is assumed that the numeric information resources of scientists must be like those of the social science librarian, and that both share a common understanding of the world of math and statistics.

Beyond these perceived similarities, there are some real links between science and social science data collections. Some scientific datasets are structured like rectangular social science data files. The activity of subsetting the data is then similar. Further, scientific datasets often are received as ASCII flat files which must be described by some kind of a programming language and converted into a standard file format prior to being used. Documentation describing the data and methodology usually accompanies the dataset. While the traditional tools of libraries designed for organizing text and citations into manipulable databases cannot do the same for numbers, certainly the tools of the social scientist can. The statistical packages used by the social scientist can retrieve and display numeric information from datasets. These tools do in fact link social scientists and some scientists: SAS and SPSS are used by both groups. In addition, scientific and social science data users are often linked because they make shared use of the computing center staff, and statistical consultants who support statistical software. When data librarians start to provide GIS data, they also bring themselves closer to the sciences. Geographers have always bridged the gap between the social sciences and sciences by their use of both social science and earth science data. Many GIS data collections which provide valuable boundary information also come with earth science data, and as a result become a common resource for both social scientists and earth scientists. Geographers also want access to images from remote sensing which once acquired will also be used by other researchers from the physical sciences.

What Are the Differences?

These similarities and linkages between the worlds of science and social science data may spur data librarians to investigate scientific data more closely. This investigation should lead to the recognition of some significant differences between the two types of data. These differences fall into three main groups: first, there are many unique problems caused by the structure of scientific data; second, a different body of knowledge is required; and finally, there are unique needs for visualization of scientific data. Data Structures Scientific data are very often structured differently than social science data. The numbers in the files, the structures they constitute, and the types of the files can all be very different than most social science data files. Most social science data files are delivered in ASCII form with a series of lines each constituting a record or part of a record and are made up of numbers which are measures of different

attributes. These numbers are usually integers. This is in contrast with the numbers in a scientific data file. Those numbers are described in terms of a much more complex classification scheme. That scheme has descended from computer programming since it has always been very common for scientists to write their own analytical software. While in social science data, numbers may have a dollar or time format, or a certain number of decimal points, in scientific data numbers are described as floating point numbers, short integers, long integers, IEEE format floating point numbers, etc. This categorization reflects the common use of binary files to store scientific data, and the many different ways that computers can store numbers in binary form. Choosing the right number type is important to insure correct results, quick processing, and efficient storage of data. In the files in which scientific data is supplied, these individual numbers are also not always structured as a series of records made up of a number of measured variables each stored as one number. Instead, the numbers may describe a point, scalar, or vector measurement and so an individual variable may include one or several numbers. In addition, lines of numbers in the file may not be series of records, but may instead represent measurements over a grid, and as a result, the number at each row and column position represents a measurement at a different place in the grid. In effect, data for all of the lines in a grid constitute one record. In the terminology of scientists, most social science data would be described as one dimensional. Each measurement is usually made once for each respondent.

In the sciences, where measurements are repeatedly taken for some entity which has length and width, as is usually the case for physical phenomena, data can become multidimensional. Three dimensional data is common for physical objects. Four dimensional data is not uncommon where measurements occur over time. Further, data of these varying forms are supplied to users in a number of file formats unique to scientific data. There are flat ASCII files available, but other file types are as or more common.

Crystallographers use data in the Cambridge Crystallographic File Format. Atmospheric scientists often use data in the National Center for Atmospheric Research's NetCDF format. And, NASA has its own file format, the Common Data Format, CDF. With these kind of specialized file types, specialized software packages are needed. SAS and SPSS do well enough with flat ASCII files, but they can't read many specialized scientific file types. Typical social science statistical packages also do not do well at representing the specialized structural properties of scientific data, e.g. 3 dimensional collections of vector values. The types of files created by these packages also do not often have sufficient mechanisms for storing the metadata which needs to accompany scientific files. CDF, NetCDF, and other file types can store information about the measurement units of the data, the names of the data's dimensions, the length of those dimensions, and calibration factors, among other forms of

metadata.

Methodology/Subject Expertise

A social science data librarian can have skills, tools, and data resources that are used across a number of social science disciplines. In the sciences these vary from one discipline to the next: a knowledge of chemical data doesn't help much when dealing with high energy physics data, or models of the magneto-sphere.

The social sciences seem more like one of the subdisciplines of the sciences than an equal to the sciences as a whole. Like chemistry and its subdivisions, there is a substantial body of common methods. Scientific data is also not as accessible to the lay person as a social survey. It is easy to understand the questions asked in a survey and the range of responses. It is not as easy to understand scientific tests and their outcomes. Such tests are often referred to by their technical name and assume a knowledge of the relevant field.

Graphics and Visualization Graphical representation or visualization of data is also different in the scientific world. Scientific data is much more frequently given graphic form as part of the research process than has been true for social science data. Models of fluids can only be appreciated via graphic representation. Molecules need to have their structure drawn to be easily differentiated. With most forms of scientific data, it is important to know how to give the data graphic form. As part of exploratory data analysis, and other analytical methods, social scientists do create graphic representations of their data. However, very often it is possible for analysts to use frequency tables, statistical calculations, and cross-tabulations which never produce graphical output. With many forms of scientific data, the object being studied can be unintelligible without graphic representation.

Images are very important to scientific research in additional ways. Much scientific data consists of images produced by cameras, or pseudo-images produced by other forms of sensing like x-rays. These images require specialized tools which can facilitate their presentation and conversion between different file formats. These tools also must provide image processing capabilities for filtering, smoothing, and otherwise enhancing the image. Both photographic images and images produced by other means will need to receive this treatment. Images are not something used very extensively in the social sciences. Graphical rendering of scientific information not only requires an appreciation of the techniques used in a particular discipline. It also requires attention to be paid to the factor of human perception. Different colors can affect the perception of the size of objects, or the attention paid to parts of a graphic. The way that images are drawn or processed is a methodological issue for the sciences.

How Do You Cope With the Differences?

Motivated by user need, or intrigued by the challenge presented by these differences, a data librarian may choose to determine what is required to provide access to scientific data similar to that provided for social science data. The differences between scientific data and social science data are significant but not insurmountable. The required subject expertise can be borrowed from a number of sources and encapsulated in data management systems. Scientific visualization software can also be obtained to handle data management and graphical rendering tasks.

Scientific Visualization Software

It will be necessary to invest in new software tools. Software must be purchased which can read and write the required file formats. The best software tools available for reading scientific data files are scientific visualization packages. Scientific visualization tools also incorporate image processing and rendering routines. These allow the creation of graphical representation of data, the display of images, and analytical processing of both data and images. To cope with the problem of investing in new software tools, cooperative efforts should be made with the computing center, scientific departments, or other universities. Software can be cheaper if it has been site licensed, a department already has a server with the software on which data can be placed, or if the costs can be shared among a number of institutions. Consulting assistance, and graduate student workers will also more likely be available for software packages already in use locally. Subject Expertise Disciplinary experts need to be found to assess scientific data collections, and software tools, and to respond to public queries about the data collected. These may be library staff, like the chemistry librarian, or they can be staff of a science department, such as the science department's computer guru. Existing data library staff could be used if they spent the time to obtain the necessary disciplinary expertise, but this would be a tremendous cost to be paid by the data library. Without additional outside assistance, no extensive scientific data service program is possible. Once outside expertise is obtained, the role of the data librarian must be to serve as the expert on data management in its most general sense. Resolution to the issues already well understood by the data librarian, such as what file formats are used and what software is required need to be answered. The data librarian should supply the list of data management questions which need answers and ensure that the disciplinary expert provides the needed information. Because subject expertise often comes from outside the data library it may not be available on a regular basis. To compensate for this, it is necessary to encapsulate that expertise in the form of software. The software commands and procedures for file handling, visualization, and data extraction which are most commonly used by experts need to be incorporated into a menu-driven data management system. The resulting program would resemble recently developed programs for social science data extraction using a WWW interface. With these extraction programs, typical steps and statistical software package commands are auto-

matically executed for the user of a WWW browser. The user is only required to make a selection from a limited number of actions appropriate for the dataset, and who need not understand the syntax of that software package or the physical format of the data. Thus, automation makes it possible to give inexperienced staff or librarians, including data librarians, an easy to use tool which doesn't require them to confront the complete set of program manuals, or the issue of the proper syntax, every time they want to access the dataset. These tools can then also be made remotely available in the science library.

Conclusion

The social science data librarian has the knowledge necessary to understand in general terms what it takes to provide access to scientific data, if not the specific skills or tools needed to implement this access. We know the questions that need to be asked and the kind of experts who can answer them. Issues of data structures, software tools, relative importance of datasets, and disciplinary practice are common to both areas. We can act as managers of data services for both forms of data. There are many potential beneficiaries of such an effort to create broader numeric data services. The library may gain additional support and respect from the scientific community. Data libraries, many of which are now under tremendous financial pressures, may gain new sources of funding. The user of scientific data will gain in many of the same ways that social scientists have gained from centralized provision of data. Data will not always be locked up in one faculty members office, or in the hands of another unapproachable department. Data collections that might not have been acquired because their user communities were divided can be purchased via cooperative cross department efforts. Producers of scientific data will have expanded facilities through which to share data with their colleagues and students.

*, Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

Building an Archive of U.S. Census Related Data Products

by John Blodgett*
Urban Information Center
University of Missouri St. Louis

The consortium for International Earth Sciences Information Network (CIESIN), in conjunction with the Urban Information Center at the University of Missouri St. Louis has established a public archive of United States census data. The data are available via the Internet using FTP and/or a WWW browser. Currently the archive contains map boundary files for all common geographic units used in the census (including census blocks, block groups, tracts, counties, etc.) in a standard ascii (BNA) format. Data extracted from the 1990 decennial census Summary Tape File 3 (STF3) are provided in a format that makes it easy to link with the boundary files to create thematic maps with widely available GIS software such as Atlas*GIS, ArcView and MapInfo. These data files are organized by state and geographic level. Using the archive researchers should be able to readily retrieve files that will allow them to analyze and/or map data for areas as small as blocks or block groups, for anywhere in the U.S.. This paper discusses the content of the archive as well as describing some of the details of how and why it was constructed.

Archive Overview

The archive of census-related data products (CRDP) is a subset of the data archive maintained by SEDAC (the Socio-Economic Data Application Center) at CIESIN. It is a joint venture with the Urban Information Center at the University of Missouri St. Louis. The UIC develops programs (in SAS) for accessing the raw census files and creating data products. CIESIN then adapts these programs and creates the "assembly line" applications (using SAS and/or Unix shell scripts) that actually generate all the specific data files that are the data archive. The CRDP archive currently consists of over 11,000 files (some of which are mini-archives created with the zip software, which when unzipped may turn into several data files) and over 4 gigabytes of data (compressed — multiply by 3.3 to get uncompressed size.) It contains the following basic kinds of files:

- Mapping boundary files in ascii (BNA) format. These files can be used in desktop mapping packages to create maps of various types of geography such as counties, census tracts or blocks.

- Extracts of 1990 census files with over 200 variables such as total population, total households, median household income, age and race distributions, etc. Intended to match up with the boundary files to make it easy to create thematic maps

showing these data values. (The boundary and census extract files are stored in almost identically structured parallel directories within the archive.)

- Street intersections files. One file per county in the U.S. with one record for each pair of intersecting street features (as found in the TIGER files) with latitude, longitude coordinates.

- ZIP equivalency files. One file per state in the U.S. with one record for every populated 1990 census block with the latitude, longitude coordinates of an internal point and a long list of other geographies associated with the block, including ZIP code. These files can be used as a powerful tool to build geographic correspondence tables.

- A special collection of enhanced county to county migration files showing the characteristics of persons moving within the U.S. between 1985 and 1990. The Census Bureau's product code for this data collection is STP28.

How to Access

The archive can be accessed via anonymous FTP at <ftp.ciesin.org> by going to the directory `/pub/census`. You can also use a web browser to go to the CIESIN demographic data home page (<http://www.ciesin.org/datasets/usdemog-home.html>) and then selecting the hypertext link to `/pub/census`. The latter method is preferred for your first visit since it gives a more user-friendly interface and can help you get a better overview of how the archive is organized. Experienced users may find that a direct FTP connection is more efficient, especially for retrieving files in quantity. Also, the web browser FTP function transmits all files in binary mode only; this is a slight problem for text files transmitted to PC environments since you do not get the extra carriage-returns added between records as you do with regular FTP in text mode.

Most of the data in the archive (everything but small text metafiles) is in zip-compressed format. You will need to have software to unzip at your end (pkunzip works fine); the archive has a special directory (`/pub/census/src`) with a selection of binary versions of zip/unzip that you can retrieve and use. Having access to and familiarity with using FTP and the unzip program is critical for anyone wanting to access the archive for more than just a few files.

The Directory Structure

The archive directory structure is fairly simple. Under the "root" directory (/pub/census/) are two relevant subdirectories:

- src, containing the zip/unzip software and

- usa, containing all the data for the U.S. (whenever we refer to a directory it will be assumed to be a subdirectory of /pub/census/usa.)

There are four relevant subdirectories of usa:

- stf: containing the census extract files (derived from the Summary Tape File 3 census tabulations, hence the name "stf")

- stp: containing the county to county migration files (derived from stp28, the Census Bureau's code name for the county to county special tabulation product).

- tiger: containing the mapping boundary files and the street intersections files.

- zipaq: the ZIP equivalency files.

(There are actually others, but we are restricting our discussion of the archive to these four primary subdirectories.)

Each of these major subdirectories has its own substructure to help organize the data within them. An important thing to keep in mind is that the stf and tiger directories have parallel structures so that when you navigate your way to retrieve a boundary file from tiger you should be able to repeat the path through the stf structure to obtain the corresponding attribute file (demographic extract).

The TIGER Directory

This is by far the largest and for many the most important portion of the CRDP archive. It includes a series of doc and txt files to provide various documentation and geographic code lists that can help you find things. It has its own "lcode" subdirectory where you can find some code (SAS programs) that may be retrieved and used to process the data in SAS. For example, one of the files in tiger/lcode is called bna2sas.sas and is a program to convert bna format boundary files back into a format that the SAS/GRAPH GMAP procedure (and, some day, SAS/GIS) can use. But the essence of the tiger directory is its collection of over 50 subdirectories corresponding to the states and territories. To access data for Missouri you use:

```
cd /pub/census/usa/tiger/mo .
```

Each of the tiger/ss subdirectories has an identical structure of four subdirectories. (Note: you really do not have to

"learn" all this — as you issue cd commands in FTP to "descend" the directory tree, helpful readme/message files are automatically echoed to the screen to guide you.) Most of the data is contained within the bna_st subdirectory; it is where boundary files for all the substate geographic levels are found except for the block level files. The bnablk subdirectory is where you can find the set of block boundary files, one file per county in the state. The csvisd (Comma Separated Values, InterSections Data) subdirectory is where you can find the street intersections data. A fourth subdirectory may be present for some states and may or may not be empty. It is called xptsdl and contains a set of SAS transport format files that represent the TIGER line files for each county. It is beyond the scope of this paper to deal with these files; suffice it to say that they would be of interest to a very small audience of persons who wanted to use SAS to access street level data from TIGER.

Descending down one more level to the bna_st subdirectory may be one of the more daunting steps of your journey through the archive. When you type "ls" to get a listing of this directory, what you see will vary from state to state, but will have a common structure. The explanations for the file names here are provided in the .message file from the previous level. You need to look at those notes very carefully so that when you look at the /tiger/mo/bna_st directory listing you will understand that, for example:

ap29.zip is the A-pumas file for the state (29=mo FIPS code).

bg291740.zip is the block group file for metro area (MSA) 1740 in Mo.

You could get the file tiger/msacodes.txt to have a listing of all the MSA codes in the U.S. that you'll need. Another way to know the codes is to first get the stf data for these areas; when you cd to stf/mo you'll see a .message file that tells you the names to put with the metro area codes. You would then know that 3760 is the code for Kansas City and that would let you deduce that t293760.zip (back in the tiger/mo/bna_st/mo directory) is the tract boundary file for the Missouri portion of the Kansas City MSA. What you'll note is that there is one bg (block group) and t (tract) boundary file for each metro area in the state plus one for the pseudo-MSA 9999 (remainder of state). The other files are statewide: bp29=B PUMA's, c29=counties, fm29=FIPS MCD's, fp29=FIPS places (cities), m8_29=1980 MCD's, and t8_29=1980 tracts (where defined).

The TIGER line files from which these mapping files were created were supposed to be "perfect" with respect to what is called "topological closure" — which really just means you can take all the line segments in them and create chains of segments that form the boundary of any geographic area and those line segments should form a complete, closed polygon. But perfection is difficult to achieve and our efforts in trying to achieve a 100% complete mapping data base has not yet

been attained; we are at 99.8% and the chaining rate tends to be best at the smallest levels. The file tgr29all.zip (in the MO directory) is a report file that contains information on the chaining process for this state.

If you select either the bnabl or the cvisd subdirectories from the tiger/mo subdirectory, then an "ls" will yield a similar list of 115 files corresponding to the 115 counties in the state. Note that with it is very easy with FTP to go to one of these subdirectories and enter the command

```
mget *
```

to retrieve all the files in that subdirectory. But be careful before you do this as it may involve transmitting more data than you will have room for on your disk. Also keep in mind that these are all compressed files and that they will typically grow to be 3 to 5 times larger when they are unzipped. You need to consider strategies for how you are going to process this much data and where you are going to store the results. And as long as we are saying be careful, the most frequent error users make when accessing the archive is to forget to issue a "binary" command before getting ".zip" files. You should only have binary turned off when retrieving text files.

The STF Directory

This directory has a structure which very closely parallels the tiger directory. It has 51 state-level subdirectories (states and DC only) and a "us" subdirectory that is rather special. It also contains a number of explanatory text files (such as "xtabs3" and "fafvar") that should be retrieved and examined carefully. There is also a 0code subdirectory with some useful programs for processing the files you get. It includes a SAS program that will read the pair of .csv files as downloaded from the archive and turn them back into the SAS datasets from which they were originally derived. The us subdirectory, a very recent enhancement, allows you to access all data of a certain type (e.g. tract level data or ZIP level data) in a single subdirectory. This directory is done by creating a series of links to the actual data files which physically reside in the 51 state-level subdirectories.

As mentioned in the discussion of the tiger subdirectory, one of the nice things about working in the stf directory is that when you cd to a state subdirectory, a .message file is displayed to the screen which is an annotated table of contents. For example when I type "cd stf/mo" I see:

CONTENTS of mo

File	# recs	Description
county<a/b>	115	MO counties
mcd<a/b>	1,367	MO mcd/ccd's
tr1740<a/b>	28	tract/BNA's COLUMBIA, MO
tr3710<a/b>	31	tract/BNA's JOPLIN, MO
.....		
place<a/b>	1,014	MO places (cities)
place<a/b>	960	MO places (cities) <sic>

```
metro<a/b>      5    MO (c)msa's
zip1740<a/b>    9    5-dig ZIPs (1991) COLUMBIA,
MO
...
zip9999<a/b>    714   5-dig ZIPS (1991) Non-metro
rmdr of MO
```

What you are reading here is a file that was actually written by the same program that wrote all the files that it is summarizing. The program did not know they were going to be zipped so it does not show the ".zip" extension that is actually present on each of the files. The "<a/b>" notation indicates that all these lines represent a matched pair of files; what we really have are two files, countya.zip and countyb.zip, that contain the data for the 115 Missouri counties. (When you retrieve these files and unzip them what you have inside are countya.csv and countyb.csv.)

You may note similarities with the tiger subdirectories but important differences. The names are not identical; c29.zip is used for the BNA file in the tiger subdirectory, county is used here. The breaking up of the tract and block group data by metro area is the same, although "tr" is used here in naming the tract files, just "r" for the mapping files. The metro files here have no match in the tiger mapping files. Neither do the ZIP data files. This is an important point: the archive does not have files for mapping by ZIP, but it does have census data by ZIP.

Processing Retrieved Data

We have talked about the structure and content of the data archive. Now we want to discuss how a user might process the information in the archive once it has been transferred back to their computing site. Our primary emphasis will be on mapping applications within a GIS (geographic information system).

Obviously, the first step in processing almost any of the files retrieved from the archive is to de-compress it using unzip or pkunzip. If you do not have one of these programs you can download one for your operating system from /pub/census/src. If there is no module there for your system you may have a problem. MVS is one platform for which I know of no public software that will "unzip" data sets.

BNA files

Mapping boundary files are in "bna" format. This is a relatively simple ascii format that is used by Strategic Mapping, Inc. (the producers of Atlas*GIS/PRO software) as the standard format for importing/exporting their mapping files. In order to use one of these .bna files you are first going to have to convert it to a format that is directly usable by whatever software you plan to use to produce your maps. If you are using one of the SMI products (Atlas*GIS or Atlas*PRO for DOS or windows) then you need to also obtain their IE (import/export) utility program. The BNA files in the archive have been created with the capabilities of

the ie utility in mind. If you go in to edit/browse a .bna file you will see something that looks like this:

```
"211739802.00106A","106A","1211739802.00106A","blk",4
-83.940732, 38.067433 -83.942302, 38.066794 -83.940591,
38.067890 -83.940732, 38.067433
```

These 3 lines represent one polygon. The first line is the header record which identifies the feature with a series of 4 identifier fields and then gives the count of points to follow. The next two lines contain the latitude longitude coordinates of the points which describe the shape of the polygon. This particular example (borrowed from the example in Henk Meij's APDU paper) is for a census block in county 21117 (somewhere in Kentucky.) It is block 106A in tract 9802.00 (BNA 9802.00 to be technically correct, but we hate to use the Census Bureau's correct terminology for these tract equivalents since "BNA" already has one meaning in these discussions). The first quoted field on the header record is called "polid" and is a unique string consisting of the FIPS state (21) and county (173) codes, the tract/bna code (9802.00) and then the block code. (When and if we ever put block level data in the archive you can be sure that the first field on the corresponding data file for this block will have a value identical to this polid field for ease of linking the data to the polygon.) The second field on the header, "106A" is called "name2" and gets stored as the secondary name field when it imports the file. This field is typically used as a name field that is often used as a polygon label for identifying the polygon on plots. The 3rd field on the header, "name3", has a value identical to the first but with a digit "1" pre-appended. This gets into some technical stuff about Atlas*GIS, including the differences between the DOS and Windows versions. In the Windows version, this field is used as the unique (across all layers in a file, not just for this layer) internal identifier. In DOS that identifier was generated for us and we did not have to worry about it, but in Windows we have to create our own. Finally, the 4th field on the header line has the value "blks". This is used as a layer identifier value. It says to store this polygon in a layer to be named blks. If you wanted to create a mapping file with counties, tract and blocks, all together in one mapping file, you could concatenate your 3 bna files for the different levels and import them as a single file. The result would be a single geographic file with 3 layers; the names of those layers would be based on the value that appear as the 4th field of the header records in these files. You can always easily rename them after the geographic file is created. Atlas users should note that concatenating files in this way will make importing coverages much faster because Atlas will only have to index the files once. When you add layers one at a time, the program has to keep going back and regenerating the indices which can be very time consuming for large files.

Note that the last field on the header is a numeric node count with a value of 4. But if you look at the data on the next two

lines you'll notice that the last coordinate pair matches the first so what we really have is a 3-point polygon; this block is a triangle.

CSV Files

"CSV", for "comma separated values", is the standard extension for these files in the Windows environment. All stf3 extracts are stored in this format, with the first record of each file containing the field names for the variables instead of data values. This format was specifically designed for ease of import into the ATLAS*GIS/PRO software packages. Importing these data files and linking them with the corresponding regions file is very simple with the DOS version of the A*G products. There are some problems with the Windows version which are associated with the new "standard" for processing CSV files which are incompatible with the DOS standard. The problem has to do with the software changing character fields which do not contain non-numeric characters to numerics, even though their values are enclosed in quotes. This same problem occurs when importing the values into Excel.

CSV files can be easily read by SAS using the appropriate infile statement options. There are a number of sample programs in the archive which illustrate this. The x3csvsas.sas program in the stf/ocode directory should be retrieved and used by anyone wanting to process these data with SAS. (In some cases these data are also available in the form of SAS export-format files in stf/xphtix subdirectories — but not for all states.)

It is beyond the scope of this paper to go into detailed descriptions of the kinds of data analysis applications that can be performed using these data. There is a good chance, however, that examples of such applications will be added to the archive in the future.

Possible Future Directions

There are a number of areas in which the archive is looking to expand in the near future. Among the kinds of data that have been discussed for possible inclusion are:

- 1980 STF3 data comparable to the 1990 data.
- 1990 STF4 data.
- 1990 STF3 data allocated to PUMA geography.
- 1990 place-of-work data from the CTPP package.

In addition to more data, there is a strong possibility of new processing options for persons accessing CIESIN's WWW site. Custom on-the-fly data extracts and using interactive maps to allow users to choose geography (by pointing to the map) and to display results are among the items that are being considered. An experimental mapping application (called "YAVEMS", for "yet another very experimental map server") is already available as a prototype of the sort of things to come.

And finally, there are plans to prepare a detailed User's Guide to help people find and utilize the considerable but occasionally confusing contents of the archive.

Credits

Most of the credit for the CIESIN archive goes to Hendrik Meij of CIESIN. While we in Missouri wrote some programs Henk is the one who had the vision of putting those programs to work and creating this public archive where people could benefit from the resulting data products. We began building the archive prototype in 1992 as our "hobby project". It received no formal institutional support until it was adopted by the SEDAC people in 1994. Clearly, the people at SEDAC/CIESIN deserve considerable credit for the support they have provided the project since then.

The Missouri State Census Data Center in particular and the Census Bureau's nationwide State Data Center program in general provided the environment in which we were able to develop our software tools for processing the census files. Since we were an SDC agency the Bureau provided much of the data used in the archive at no charge. And, of course, none of this would have been possible without the excellent raw materials we had to work with, virtually all of which was provided by the U.S. Census Bureau. They have also been very supportive whenever we have had to go to them with questions about their data products.

Much credit goes to the data archive at Lawrence Berkeley Labs (LBL) where we were able to go and obtain (via FTP over the internet) many of the raw STF files and some of the TIGER line files. We also had files provided by the New Jersey State Data Center (at Princeton) and the California SDC. We have also received much valuable feedback, constructive criticism and inspiration from many people all over the country — most of them subscribers to the SASPAC-L listserver. I would name names but the list would be too long and I'd surely leave some out.

And, last but not least, much of the real credit for the success of this project goes to the Internet itself. Not only is the archive a concept that would not make sense without the internet, it is one that was created on and because of the Internet. It was the net that brought together two people with a shared interest in the application, one from Missouri, the other from Pennsylvania - who otherwise would probably have never heard of each other, much less been able to collaborate on such a technical, data-resource intensive undertaking as the building of this archive. The Internet was the critical tool that allowed quick sharing of ideas, of source code, of sample files, of user comments, of sample outputs, and finally of the resulting products. And when we thought we could never afford to obtain all the STF3 data on tape and convert it, we were able to overcome that barrier by using the Internet: LBL supplied the raw data as .dbf files on from their cd-rom based file server and a person named Richard Hockey in Australia was able to make us aware of his SAS

code to convert .dbf files to the SAS format we needed. That tool came to us via the SAS-L listserv and then via FTP to fetch the advertised code from its home library in Australia.

* Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

Developing a Scottish Migration Monitor: a co-operative approach

by Alison McCleery* & Emma Forster Department of
Economics, Napier University
Heather Ewington & Peter Burnhill Edinburgh
University Data Library, University of Edinburgh

Abstract

This paper documents progress on setting up a Scottish Household Migration Monitor and, in so doing, traces the path of a learning curve which operated from the inception of what came to be known as the Migration and Housing Choice (Scotland) Survey through to the collection, processing, analysis and interpretation of the data it generated. Hosted by Edinburgh University Data Library, such a Monitor will be available to the academic community, and to public and private sector housing and planning agencies.

An innovative feature of the Monitor is the inclusion of information on migration motivation as well as the numbers and patterns themselves. The inclusion of such qualitative - as distinct from quantitative - data posed questions about how best to handle both data types. It is an important issue to resolve since better informed decision-making is made possible as a result of documenting and understanding the decision to migrate and how it varies in space and time. For example, urban and regional planning authorities will be able to estimate housing demand and future housing land requirements more accurately.

Introduction

The Migration and Housing Choice (Scotland) Survey was conducted in the early nineties by researchers from two Scottish Universities; Strathclyde and Napier. The purpose of the survey was to discover the *intentionality*, namely, the motivation behind household migration patterns, to use the data for academic research and to inform decision making by urban and regional planning agencies in both the public and private sector. We regard the work done to date as a large pilot study for what may become an ongoing Scottish Household Migration Monitor.

This paper contains descriptions of the following:

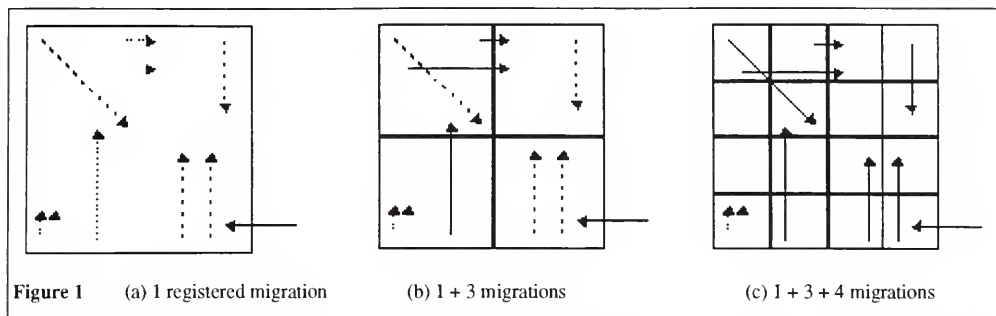
- * context of the survey
- * conduct of the survey
- * data handling issues
- * the role of Edinburgh University Data Library
- * assessment of strategy and future plans

Context of the survey

Of the three components of population change, namely fertility, mortality and migration, it is migration which is the most significant, whether for facilities planners in the public sector or for market researchers in the private sector. With the decline of both fertility and mortality in recent decades, migration has become the primary process responsible for changes in population numbers and composition, with that importance magnified at the local scale (Champion, 1993). However, migration researchers may experience greater problems than researchers into fertility and mortality when they attempt to use existing data sources relating to their respective fields of study. Official data agencies release data at levels of aggregation which, on the one hand, protect confidentiality, but on the other hand, hide much migration activity. The following three diagrams illustrate how a move 'on the ground' becomes recognised as a migration only after it crosses a specified boundary, ie, is externalised.

Because the crossing of a boundary has become the defining feature of a migration, the importance of distance travelled is underplayed. For example, the moves depicted by the two horizontal arrows in Figure 1(b) would both register as migrations, regardless of their markedly different lengths. Yet the diagonal arrow in the same quadrant, although the same length as the longer of the two horizontal arrows, would not count as a migration at all, even although it is more than twice the length of the shorter horizontal arrow. Only at the level of disaggregation shown in Figure 1(c) does the long diagonal arrow register as a migration. However, even at this level the two short-distance moves in the bottom left quadrant do not count as migrations. Thus, not only is some migration information concealed altogether, but also distance information is lost, which is

significant because short-distance moves indicating *residential mobility* are more important numerically than long-distance migration and further, that the motivation behind each type of move is different.



The solution suggested by Forbes and McCleery (1991) is, in an ideal world, to attach a specific geographic reference to the data at the point of collection which would then permit analysis of all moves and provide the flexibility to aggregate the data to whatever level, with due regard to confidentiality.

Migration researchers in Scotland are more fortunate than most in having access to the **Register of Sasines**¹. The Sasines Register is a land register, unique to Scotland, which records every private property transaction and therefore allows every moving household in the private sector to be identified. The information contained in each record is the name and address of the owner of the present property, a previous address (of questionable accuracy - see below) and the price of the sale. Used judiciously, the Register of Sasines provides a reasonably accurate record of the origin and destination of moving households by precise postal address. Tracking of movements is theoretically possible at the lowest geographic level, namely the household. However, in practice, although the present address is accurate because it is recorded from legal documents relating to the sale of that property, the previous address information is collected mainly for administrative reasons. It may refer to the previous long-term address, but equally could refer to temporary accommodation occupied prior to the move or might even be the new address if there was a delay in completing the legal documentation. As McCleery (1980) emphasises, as a source of migration data, the Register of Sasines has to be used with care.

Furthermore, even if it is possible to produce accurate patterns of moves on the map, it should be recognised that these are merely the visible trace of an invisible, complex, increasingly segmented and largely unexplored household decision. It is the accumulated decisions of all the moving households which drive the migration streams. Without some knowledge of these decision processes, and how they may vary from place to place, from time to time, and from household to household, it is impossible to estimate future volumes and directions of moves. Thus is introduced an unknown factor in estimates of housing demand, and consequently housing land requirements are difficult to assess.

Although to date, some research has been carried out investigating movers in Scotland (Garner, 1980; Forbes, 1989), there has yet to be a comprehensive survey of all tenures at the national level. Forbes' paper examined migration patterns in Scotland's largest city, Glasgow, but was more innovative than either descriptive or analytical in that it used pre-existing data to explore the feasibility of an experimental Migration Monitor. Now this present paper reports on the broadening of the initial work into this area. In so doing, it traces the path of a learning curve which operated from the inception of what came to be known as the Migration and Housing (Scotland) Survey through to the collection, processing, analysis and interpretation of the data it generated.

Conduct of the Survey

The survey's two principal investigators were Jean Forbes of the Centre for Planning at Strathclyde University in Glasgow and Alison McCleery of the Department of Social Sciences at Napier University in Edinburgh. Building on her previous work, the current project was initially designed by Forbes to explore the various elements in the decision to move within the West of Scotland. The subsequent involvement of McCleery allowed coverage to be extended to include the whole of mainland Scotland. The Data Library, located at the neighbouring University of Edinburgh, was brought in at a later stage to introduce a greater element of economy and efficiency to the questionnaire design and subsequent data processing.

The survey data was gathered by means of a postal questionnaire mailed to a 25% sample of households having moved between January and October 1990, identified from a computerised version of the Sasines Register held by the Land Value Information Unit at a fourth Scottish University, Paisley University. The unit is a commercial concern and deals with requests for data from organisations such as chartered surveyors, builders, local authorities and researchers. Computerised housing sales records for the whole of Scotland are available from 1989 onwards and, for some parts of the country, from 1979.

Mailing of the questionnaire was made possible with the co-operation of a number of local government planning departments² whom the investigators had succeeded in interesting in the project. The Royal Mail (Post Office) also helped because they had recognised that the survey would collect useful information about public awareness of the postcode³; the questionnaire asked respondents to provide both their previous and present postcodes.

The design of the questionnaire was informed in two ways; firstly, it was influenced by the elements in an *a priori* model of the process of decision-making developed by Forbes (1989) and secondly by elements related to housing supply and local environmental quality proposed by the collaborating planning departments. In essence, the questions were divided into 4 types:

1. question initiating open-ended answer as text
where was your previous house?
2. question with pre-defined answer categories, only one answer per question
what type of house?
Detached [] Semi-detached [] Terraced [] Flat []*
**Apartment*
3. variation on the above, essentially one question, but with each possible factor presented as a separate question with only one answer per question, either yes or no
what factors influenced your decision to move? job transfer []
4. question inviting multiple, open-ended answers as text
which other localities did you consider?

During the course of the project, the questionnaire was enhanced as a result of further consultation with the planning authorities from whom we sought help with distribution of the questionnaires. Variations between versions were minor, consisting mainly of additional questions, e.g. including a question about car ownership and a modification allowing respondents to specify *other* reasons for leaving the previous house and choosing the present house, other than the pre-defined reasons given on the questionnaire. The inclusion of the latter option allowed respondents to provide information in their own words about their motivation for moving. The challenge later was how best to translate these subjective snippets of information into objective *data*. Apart from these additions, consistency was maintained to allow comparability of the core questions.

As expected with a postal survey, the overall response rate fell short of 50%, although this varied decidedly from District to District. The resultant 10,006 cases represented about 1 in 10 of all private sector residential movers for the period. A proportion of Strathclyde respondents volunteering their telephone numbers was also followed up with a more detailed survey of search patterns.

Data handling issues and the role of Edinburgh University Data Library

Data collection

The survey data were collected and processed at different times by different institutions. A pilot study was conducted first on one area of Strathclyde Region, the sprawling, industrial city of Glasgow. The second, and larger survey covered six Regions; the rest of Strathclyde, Dumfries and Galloway, Fife, Grampian, Tayside and Central. Forbes at Strathclyde University organised data collection for both surveys and because of the large number of replies, considered it cost-effective to use an outside bureau agency to process the data. The two Regions of Fife and Highland, took on responsibility for collecting and converting their own data and, with assistance from EUDL, McCleery collected and processed data from the

remaining two Regions, Lothian and Borders.

The smaller number of the East coast replies, together with financial constraints, prompted the investigation of an in-house solution to processing the data. The advantage of this approach was that we gained greater control over the data, but we still had to ensure that it was in a format consistent with the bureau-processed data which, because handled first, was regarded initially as the standard to follow. The West coast data had been converted into rectangular numeric data files and processed as SPSS system files. Although Strathclyde University operate a VAX system and Edinburgh and Napier Universities operate UNIX, we did not envisage any problems with data transfer between the two systems.

In-house processing

On advice from EUDL, it was decided to use the database package **FileMaker Pro** to input and store the East coast results. EUDL had used the software successfully on a number of other projects and thus was able to confirm its suitability for the purpose. In particular, it was a user-friendly package which a new keyboard operator could quickly learn how to use. The following positive attributes apply:

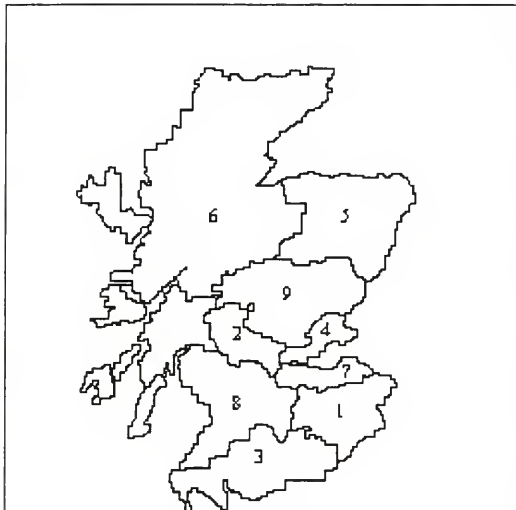


Figure 2: The administrative map of mainland Scotland

Key to Regions

- | | |
|---|-----------------------|
| 1 | Borders |
| 2 | Central |
| 3 | Dumfries and Galloway |
| 4 | Fife |
| 5 | Grampian |
| 6 | Highland |
| 7 | Lothian |
| 8 | Strathclyde |
| 9 | Tayside |

- * database definition is simple
- * the facility to construct different views of the database (layouts) allowed the construction of a data input screen which closely resembled the layout of the questionnaire - we reckoned that the similarity between the questionnaire and input screen would reduce the number of input errors
- * field display options such as checkboxes, radio buttons, pop-up menus and pre-defined lists helped the operator to input the data efficiently
- * the lookup facility enabled automatic coding of some data items - additional fields were created for each question requiring either a **yes** or **no** answer and the fields automatically filled with 0s and 1s depending on whether or not the checkbox was clicked
- * data validation mechanisms ensured that no invalid answers were input (e.g. text replies instead of numeric) and that no questionnaire could be entered into the database twice
- * the export facility provided output usable with SPSS. Unlike the bureau processed data, we exported tab-separated output files and created free as opposed to fixed data format SPSS files, thus allowing us to store variable length data items, such as the postcode

Assessment of strategy

Overall, FileMaker Pro matched expectations. Inputting data was relatively trouble-free and early results were obtained from the database prior to input and analysis within SPSS. It was useful to have these results to compare with the SPSS results to ensure that the data had transferred correctly between environments. Also, we were able to input *all* data from the questionnaires into the FileMaker Pro database, both numeric and textual, whereas not all items from the West's replies had been converted into machine-readable form.

Specifically, postcode information had not been transcribed. Although this information could still be retrieved from the questionnaires, it would be a time-consuming exercise and perhaps now, as the survey data ages, considered not worth the effort. EUDL recommended that the postcode be recorded in our data because many official statistics are postcoded, including the Population Census, and thus the code would provide the link with other important datasets. The last Census for the UK was carried out in 1991, shortly after our survey was conducted, and so we were keen to compare our results with those of the Census.

Also, there is no machine-readable version of the exact text of the answers to the open-ended questions. A coding scheme for these answers had been devised and applied to the West's data, but when it came to using it to code the East coast replies, we felt that a more detailed schema was required. And so we enhanced the original coding and applied it to our data. This has meant that some of the data has a more detailed level of classification for these variables and again, only by returning to the West's questionnaires would be able to standardise the treatment of this data item.

In sum, we think that we benefitted by building the Filemaker Pro database; we were able to stagger handling of the data, processing the less problematic data items first, producing some results and returning to the more difficult items (such as the open-ended questions) as and when time and resources permitted.

The spatial element

A fundamental, and what has proved to be an on-going operation has been processing the survey's spatial data items; ie, the locality of previous and present houses, location of places of work and other areas searched. We wanted to attach a National Grid⁴ reference to each of these items and be able to calculate, for example, the distance moved from previous to present house, how far respondents travel to work and the extent of their search area. Also, with grid referenced data we would be able to use GIS technology (e.g **MapInfo**) and plot migration movements. The planners in particular had expressed interest in maps of our findings.

However, for the purposes of this study it was decided not to record the address data from the Sasines records. Instead, having used the Sasines successfully to identify the migrating households, we included questions in the questionnaire asking in which locality the respondent previously and presently lived. At the time, there were cogent reasons for this course of action. Firstly, at the start of the pilot project, the team was not aware of an easy and quick way of grid referencing the precise postal address. Secondly, there was the problem of potential inaccuracy of the previous address recorded in the Sasines record. And thirdly, in this particular migration project, sensitive motivational migration information might not be revealed by respondents nervous in the knowledge that it was attached to their actual postal address. In retrospect, this decision was a mistake. If the survey were being repeated, we would pay more attention to the geographic detail and fully exploit the resources of the Sasines, trying to overcome the confidentiality issue in a different way.

Despite the shortcomings of our spatial data it still had to be processed. Although the first batch of replies was grid referenced using printed information sources, with the subsequent involvement of the Data Library, we learned that online grid referencing facilities were available which not only would automate a laborious task, but also provide sufficient geographic detail to produce meaningful maps.

We used the following Data Library utilities to grid reference our data:

- * **Postzon File**

The Postzon File provides 12 digit grid references to 10 metre resolution. The file is extracted from a Central Postcode Directory maintained by the Post Office and contributed to by a number of official organisations such as the Office of Population, Censuses and Surveys, the General Register Office (Scotland) and the Welsh Office. In addition to grid references, it contains information about postcodes (date of termination of a postcode) and area and country codes.

- * **Index of Placenames (IPN)**

Both datasets can be used interactively and in batch mode; the latter was the appropriate method for our application. A FileMaker export file was produced consisting of, where it existed, the postcode and placename information for each reply, together with each reply's unique reference number. The more specific match was on the postcode and so we used the Postzon file first and only if unsuccessful, tried to match the placename using the IPN. Both methods however were an improvement on a grid reference obtained from the printed sources in terms of speed of processing and the accuracy of the grid reference.

Data Linkage - methodological concerns

The original decision not to record postcode information in the data from the first two surveys was taken partly because of an initial lack of awareness of its use to link with other datasets. It also related to doubts about its use in deriving a geographic reference; in particular, there was concern about the method by which a grid reference is allocated to a postcode. The practice of the General Register Office (Scotland) has been to allocate the reference to the nearest 10 metres of the centre of the building judged by eye on the map to be the centroid of the area covered by the postcode.

As Forbes and McCleery (1991) have argued, postcode units vary widely in size and shape between different parts of the country. They are smallest in urban areas, but in sparsely populated Highland Scotland, they are large and straggling, and thus the process of allocating a centroid is a nonsense; subsequently to attach a 12 grid figure reference "lends spurious accuracy to a technique which is at the very least unsophisticated". It could be that the effects are marginal. But a more pessimistic view cannot be dismissed:

Massive amounts of public money have been directed in recent years to areas of social and/or economic 'need', such areas being defined by mapped patterns of aggregated data. Yet the need may not exist where the map says it does, and the map may fail entirely to reveal a need locality which does exist

Forbes and McCleery (1991)

This draws attention to problems of spatial comparability. In addition there is also arguably a problem of comparability through time since the Royal Mail allocates and re-allocates postcodes as the built environment evolves, that is, as buildings appear and disappear. Despite these acknowledged failings, the postcode is, as explained above, currently the best tool there is for achieving data linkage.

Data Linkage - practical applications

As things turned out we had an earlier occasion to use the linkage facility; namely, to overcome the problem of the East's unavoidably small sample. We had hoped to issue statistics for specific places, but in many cases the number of replies was too few to draw any meaningful conclusions and we were also wary of breaching confidentiality.

Although, all replies had been labelled by their District and Regional identifier, we considered that statistics produced for these large and heterogeneous areas would be not very meaningful. Fortunately, suitable sub-areas were identified by the planners using their local knowledge. They identified Community Council Areas (CCAs) which were largely homogeneous in terms of socio-economic composition and also constituted possible housing market areas. The CCAs had been defined in terms of the geographic areas used to release the Population Census data, namely the Census Output Areas, which in turn comprise one or more unit postcodes. Our task therefore was to discover in which CCA each reply was located. If we had had accurately grid referenced replies and if the CCAs been available in digitised form, then the task could easily have been performed using the GIS technique, 'point-in-polygon'. However, we were not in this fortunate position and so we had to rely on another of the Data Library's data services, UKBORDERS.

UKBORDERS is a national, online service providing digitised boundary data for standard and user-defined areas based on the geography of the Population Census. By combining data from UKBORDERS and information from the planners on the composition of the CCAs, we were able to construct a lookup table detailing the CCA and Census Output Area for every postcode unit. The final step was to match the replies against this file and create two new variables of CCA and Census Output Area. We then re-released the earlier statistics produced for the newly-created areas.

Overview

A project which started life as a rather modest example of legitimate, but by nature organic, academic enquiry in the best tradition of the ancient Scottish universities, had slowly but surely changed into a monster. It was not so much Nessie as messy! This situation arose mainly from the project's funding arrangements. In short, there were no funds available at the start of the project and it was able to progress only because of personal investment by Forbes who financed a small-scale, pilot study on Glasgow, with the hope that the rest of the country could be surveyed in time as and when funds became available.

It was fortunate that we were able to interest a number of organisations outwith the academic community in the project, who agreed to contribute either financial assistance or help in kind, e.g. posting the questionnaires. However, in return, we were obliged to consult with them about the content of the questionnaire and to provide them with basic results. The effect of having multiple contributors was both positive and negative: positive insofar as the survey would never have been carried out

without their help, negative in that control of the project became dispersed, with the danger that the initial objectives of the survey would be lost as we tried to be *all things to all men*.

It was unfortunate that EUDL was not involved at the start of the project. Their assistance with questionnaire design, data input and processing was important and much appreciated but, because of circumstances, consisted mainly of rectifying errors that might have been avoided if there had been better-informed planning at the start. Ideally, their input should have been sought at the beginning, a project plan developed and adhered to throughout the project. However, we do regard the survey as a success and are currently collaborating to produce a national dataset to be deposited with EUDL and thus be available for use by other researchers.

Dissemination through the Internet

EUDL has also drawn attention to the potential use of the Internet and the World Wide Web for both publicising the data and possibly providing access. We have created a few pages about the survey and have added them to the Data Library's World Wide Web server, *datalib*. Also, we have produced a WAIS (Wide Area Information Server) index of records about the survey questions, including details of variable and value labels, whether or not the variable exists for a particular area and how it was processed or derived. The organic growth of the dataset had given us the problem of where to record details of its changing structure, but fortunately, the Web arrived during the course of the project and offered us a possible solution to the documentation problem.

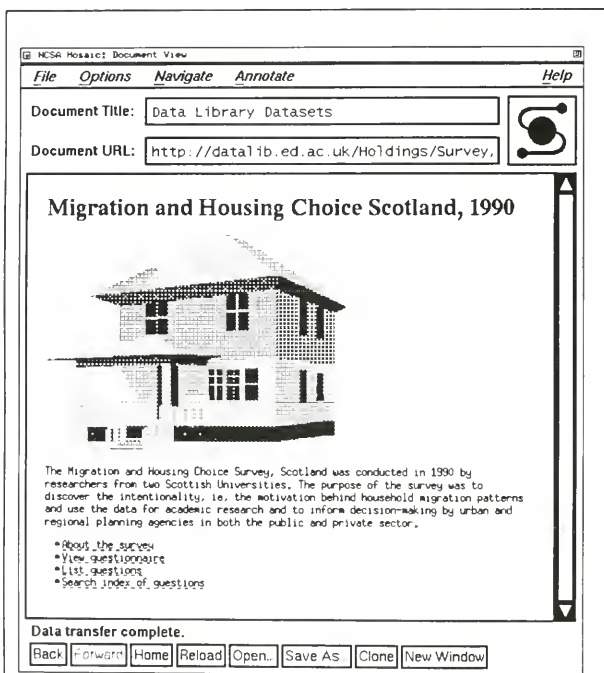


Figure 3: home page

The home page reproduced here lists the four main menu options:

- * About the survey
- * View the questionnaire
- * List questions
- * Search index of questions

We thought it useful to let potential users view the questionnaire as an aid to helping them decide whether or not the data would be of value to them. We created a gif image of the questionnaire using EUDL's scanning facilities. EUDL is involved in a number of projects involving scanning. Currently they are working with the University Library to offer Internet access to part of the Library's Special Collections by creating an index to scanned images of catalogue records. They are using a flatbed scanner controlled by the Optical Character Reading (OCR) software FormFile.

We have created a WAIS index of information about the survey variables using the indexing software, freeWAIS. Users are prompted to search the index by keywords. Additionally, they can view a question list in which each entry is a hypertext link to the appropriate index record.

In addition to the above information, we are also investigating using the Web to access the data itself directly and produce basic statistics for standard areas. We have noted other attempts to do this, including by the University of British Columbia and Statistics Canada. However, we realise that direct access to our data may require to be limited because of the confidentiality issue. For experimental purposes we intend to hardwire into the system, a selection of counts for the larger geographical areas.

Conclusion

Although subject specialists provided the

initial stimulus for this investigation and are predominantly involved in the interpretation and exploitation of results, they did not have the information-handling skills or resources necessary to create the dataset from which the results would be derived. The involvement of data management specialists, Edinburgh University Data Library, was therefore an important factor in the successful conduct of this project.

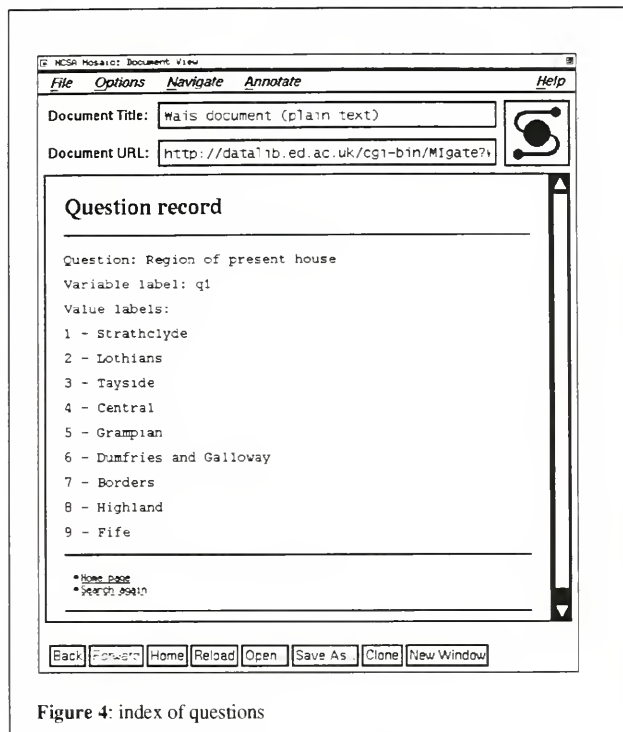


Figure 4: index of questions

Despite the many flaws, results have been produced which have indicated potential further lines of enquiry. Also, the non-academic organisations who were initially involved have reacted well to our findings. Furthermore, having been approached by a number of private sector organisations such as house builders and the government housing agency, **Scottish Homes**, we now feel confident in our ability to convince a consortium of these bodies to put together the funding for a comprehensive and efficiently organised Migration Monitor. During the course of the survey, vital lessons have been learned for the future. Should the survey be repeated in a more favourable funding environment attention should be paid to the following issues:

Organisational

- * of paramount importance is the need to agree at the outset what the survey is about; then to ask questions which elicit the correct answers, that is, answers which provide useful information

Methodological

- * we have highlighted the issue of spatial referencing, that is, problems associated with the use of the postcode as the UK standard. Frustrating though the postcode may be, nevertheless, by a process of historical accident, it has become firmly embedded in the data processing activities of UK official data collection agencies

Finally, two issues: one specific, the other more general arising from the foregoing discussion. Firstly, the present investigators have restricted the use of the data generated to supporting the activities of the organisations which contributed funding to the project. However, with the deposit of the dataset with the Data Library, the data will become more widely available and used by researchers who may potentially adopt a more challenging attitude. For example, the widely accepted use of the housing market area (such as the CCA supplied to us by the planners), could well be rejected. We have become aware that we might have been too much in the pocket of the planners. However, this is a dilemma faced by many in the UK with the shift during the 1980s from publicly-funded academic research to commercially-sponsored consultancy work. The question we may need to ask is this: what will happen if, with the wider release of the data, other users end up *biting the hand that fed us*?

Secondly, there may be a fundamental mismatch between the perspective of data professionals and the nature of academic enquiry. The former is rational, organised and has to adopt a pragmatic, problem-solving approach. The latter is organic and incremental, with a tendency to go off at tangents, some of which open up new and fruitful avenues of enquiry. For example, as we analysed the respondents own answers about why they moved house, the issue of health arose frequently; an option not given as one of the multiple-choice answers.

Issues of this type point to one of the many benefits of an organisation such as IASSIST: it serves as a forum where data professionals meet to consider all aspects of the data provider/user interface and where subject specialists are welcome to learn from and contribute to the discussions.

Appendix

Migration and Housing Choice (Scotland) Survey: some key results

These testify to the increasing complexity of motivation for migration. In other words people are citing more and different reasons for their move than previously. Respondents gave a mean of 1.48 reasons for leaving their previous address, but 2.64 reasons for choosing their new one. While this confirms the operation of a trigger for leaving the old home, it also suggests that at the level of choosing, the decision is based on multiple factors. Furthermore, it is apparent that a hard and fast distinction between pushes and pulls (i.e. *from* the origin or *to* the destination) is no longer entirely valid. For example, in response to a partner's complaints that the present home is too cramped, a person may seek promotion at work and be moved to a branch manager's position in a distant location. Was the household pushed, pulled, or a bit of both?

Perhaps this is not a particularly well-chosen example, however, since another key finding relates to the declining significance of employment as a motivational factor relative to *quality-of-life* considerations. While short-distance moves have never traditionally been predominantly job-related (as might be expected), now, apparently, this is increasingly the case also with long-distance moves. This is only partly explained by the increase in retirement-related moves associated with the changing age structure in the UK and the Western world universally. It also seems to reflect a change in the balance of importance between macro-structural and micro-behavioural determinants of - or more properly influences upon - migration. Increasing segmentation of markets for goods and services mirrors precisely the growing profusion of life choices and chances. The consequent diversification in 'lifestyle' is itself associated with an almost complete disintegration of any remaining correlation with household income.

So it is that classifications such as 'the elderly', 'the average family', 'single-person households' or even 'multi-adult households' are increasingly superficial and meaningless. Elderly people may be active, semi-independent or dependent; they may be young elderly, elderly or very elderly; they may live alone - whether as a result of being never-married or widowed - or with a spouse, sibling, offspring or companion. Their incomes, tastes and aspirations within each sub-group may vary in many ways and more. A print-out of comprehensive cross-tabulations for the elderly group alone would make serious inroads into British Columbia's forests!

Yet for all that people attempt to carve out a very specific spatial and social niche for themselves according to their particular circumstances, there exist certain quality of life variables which, taken together, identify a highly desirable local environment which, other things being equal, most people would seek after - for Europeans perhaps to live in Geneva, for Canadians, Vancouver? Nor is it even necessary to compromise on lifestyle choices for the sake of a high quality-of-life, if the two are pursued at different scales. Measurement of quality-of-life is, as suggested above, appropriate at the level of inter-urban comparison; satisfying individual lifestyle requirements is carried out at the intra-urban level. Yet the two cannot be completely divorced from each other as the following example perhaps indicates.

In the Glasgow University quality-of-life ranking of the thirty-eight largest cities in Britain, Edinburgh came out top. In our survey fewer of the respondents moving either within or to Glasgow cited liking the local environment as a reason for choosing their house than was the case in Edinburgh (56% to 64%). Evidently longstanding rivalry which has traditionally existed between Glasgow and Edinburgh is not yet dead, and the perception of Glasgow as a declining industrial city with problems of urban decay and deprivation persists. Edinburgh, by contrast, is associated with the successful financial services industry and with fine architecture and gracious parks. Moreover, our survey found that it is precisely these types of quality of life attributes which seem to be accorded higher significance by many working people than proximity to their place of employment, the latter for the present being considered a stretchable link.

Some interesting relationships also emerged between house price and distance moved and between housing density and distance moved. A higher proportion of those moving into the most expensive properties moved either longer distances or intermediate distances. This could be interpreted as a sequential process of initially a coarse-grain employment-led inter-regional move and later a fine-grain environment-led intra-regional move. The latter is less constrained by the requirement to be close to work and therefore permits a wider search area and a longer move locally than in the case of the lower paid who cannot afford the cost of commuting. This speculation is supported by the finding that a lower proportion of households migrating intermediate distances mentioned convenience to work as an influence upon their choice of destination.

Finally, it was not surprising to find that in urban areas the typical move is very short, although this is offset by a small proportion of very long extra-area moves. Less well appreciated is the opposite situation in rural areas, where the typical intra-regional move is longer, presumably because of the longer distances between localities offering a comparable level of housing choice. Thus, in the very rural district of Argyll, 62% of movers have travelled 30kms or more, as against 18% for Glasgow, Scotland's urban heartland. The short distance move figures are the reverse, with 29% for Argyll and 54% for Glasgow.

It would be possible to produce a PhD on the interpretation of the results from the Housing Choice (Scotland) Survey. Indeed this is precisely what one of the authors of this paper, Emma Forster, is currently undertaking. In so doing, she is transforming herself into that still rare breed of highly-skilled and highly-prized individual who is both a knowledgeable subject specialist and a competent data manager.

* Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

Bibliography

Champion, T. (1993). 'Introduction' in T. Champion (ed.). *Population Matters: the local dimension*. London : Paul Chapman Publishing Ltd, pp. 1-21.

CANSIM data base: Canadian socio-economic information management system [computer data]. Ottawa, Ont.: Statistics Canada [producer and distributor], [19—]. 1 data file and accompanying documentation. <http://www.datalib.ubc.ca>

CITIBASE, Fame Economic Database 1946- [computer file]. New York : FAME, 1978-1994. <http://www.datalib.ubc.ca>

Datalib: Edinburgh University Data Library's WWW server [computer data]. Edinburgh: Data Library, University of Edinburgh, 1995- <http://datalib.ed.ac.uk>

FileMaker Pro [computer file]. FileMaker Pro 2.0bv2, May 1993. Computer program. Santa Clara, California : Claris Corporation. Copyright 1988-1993 Claris Corporation.

Forbes, J. (1989). 'Migration Monitoring and Strategic Planning', in P. Congdon & P. Batey (eds.). *Advances in Regional Demography*. London & New York : Belhaven Press, pp. 41-57.

Forbes, J. & McCleery, A. (1991). The 1991 Census: spatial referencing considerations. ESRC Regional Research Laboratory for Scotland Working Paper No. 21. Edinburgh: ESRC RRL Scotland, 1991.

FormFile [computer file]. Version 1.11. Livingston, Edinburgh : Seel Ltd, 1994.

freeWAIS-1.0-sf [computer file]. Release 1.0 2/16/93. Thinking Machines, Jim Fullton, Kevin Gamiel, Jane Smith, Tung Huynh, Ulrich Pfeifer.

Gamer, C.L. (1980). *Residential Mobility in the Local Authority Housing Sector in Edinburgh 1963-73*, Unpublished PhD Thesis, University of Edinburgh.

IPN [computer file]. Computer program. Edinburgh : Edinburgh University Data Library, 1991.

MapInfo [computer file]. MapInfo Version 3.0.2. Troy, NY : MapInfo Corporation. Copyright 1985-1994 MapInfo Corporation.

McCleery, A. (1980). The Register of Sasines as a Source of Migration Data, *British Urban and Regional Information Systems Association Newsletter* 46: 16-17.

Postzon [computer file]. Edinburgh : Edinburgh University Data Library, 1991.

Rogerson, R. Findlay, A. Morris, A. Paddison, R. August (1989). In *Cities. Variations in quality of life in urban Britain*. pp. 227-233.

The Scottish National Dictionary. Edited by William Grant and David D. Murison. Edinburgh : The Scottish National

UKBORDERS: ESRC national online service for the extraction of digital boundary data [computer data]. Data Library. University of Edinburgh. <http://borders.ed.ac.uk>

Notes

1. Sasines is an old Scottish word meaning the act or procedure of giving possession of feudal property, until 1845 by the symbolical delivery of earth and stone on the property itself. Symbolic delivery has now been abolished and all Sasines are registered in the Register of Sasines which is now being converted into a computerised Land Registry.
2. At present, Scotland's local government administration is divided into two tiers; one level of 9 Regions and a second level of 56 Districts and 3 Islands Areas. Planning responsibilities are shared between Region and District, the former involved in strategic planning issues, the latter in development control matters. However, by the end of 1995, this two-tier arrangement will be replaced by one level consisting of 32 unitary authorities.
3. The primary purpose of the postcode is to assist the Post Office to deliver mail. The postcode is a combination of between five and seven letters and numbers which define four different levels of geographic unit; the Postcode Area (120 in the UK), the Postcode District (2,700), the Postcode Sector (8,900) and the Postcode Unit (1.5 million).
4. The National Grid is a reference system of squares overprinted on all Ordnance Survey maps since the 1940s. The system of breaking the country down into squares allows any place in the country to be given a unique reference code.
5. Crofting is a system of land tenure. A croft is a smallholding worked by a tenant, comprising a plot of arable land attached to a house and a right of pastorage in common with others. The sale of crofts has traditionally been governed by Crofting Law which restricted free sale. The Western Isles and Orkney and Shetland were not included in the survey because of complications associated with the unique form of housing tenure called crofting which is peculiar to these areas and which distorts the housing market there.

Public Access to Large Data Sets in a Depository Library

by Juri Stratford*, *Government Documents*
Department Shields Library University of California,
Davis

In the United States, depository libraries receive federal publications under the Depository Library Program as described under Title 44, Chapter 19 of the United States Code. Depository libraries act as custodians for federal publications in exchange for providing public access to those government publications. While the Census Bureau started experimenting with the distribution of data on CD ROM in 1985, the full scale depository distribution of CD ROMs started about 1990.

There is no single agency within the Federal government responsible for coordinating the format of the data distributed. While the Government Printing Office is the agency responsible for distributing the depository CD ROMs, GPO doesn't fulfill the role of a publisher. The format of the data files and the software to access those data files, if any, is the decision of the data producer, as is the decision whether or not to provide a given data product to depository libraries.

The early momentum behind the depository distribution centered around MS DOS. The Census Bureau began using the dBase format for their files beginning with Test Disc 2 distributed to depository libraries as an experiment in 1987. Most Census Bureau CD ROMs are still distributed in dBase format. It was easy enough for depository libraries to provide access to the CD ROMs either using programs produced by the Census Bureau when available, or in other instances using dBase.

Given the nature of the depository library program, depository libraries are reactive rather than proactive. Depository libraries receive publications based upon item selection surveys. Each item number represents a class of documents or datafiles. Depository libraries frequently do not know the file format of, and software access to, the CD ROM products before they are selected. This means that as their datafile collection takes shape, they must develop access strategies after the fact.

At the University of California, Davis, we have made specific decisions regarding the types and levels of public service that we can provide for datafiles. These public service activities include loaning the CD ROMs; making the CD ROMs available on public microcomputer stations; making the CD ROMs available via the network; and providing basic extraction of data subsets for end users. We have limited our service to the provision of files, either as created by the data producer, or custom subsets. While we use a variety of applications to produce these subsets, we do

not provide any access to analytical or statistical software, including mapping software. So, for example, while we provide mapping data on a regular basis, we do not produce maps.

The original depository CD ROMs distributions did not include front end software. Our options at that point were to work with the CD ROMs ourselves using dBase or to loan the CD ROMs. Loaning the CDs was not seriously considered at this time because very few people had access to CD ROM drives. Our solution was to work with the datafiles, create subsets, and distribute the data on floppy diskettes.

As the data producers began to develop some user-friendly front ends to their data, we began to make these CD ROMs available on public microcomputers. At present, we have five MS DOS microcomputers that can be used by the public to access depository CD ROMs. Four of these CD ROM stations are in a public area. We provide access to about sixty CD ROMs on these four public stations. The primary consideration for mounting a particular CD on these public stations is our subjective evaluation of the front-end software provided by the data producer. We mount only files with appropriate front-end software for unmediated public use on these machines.

The fifth CD ROM station is a computer in the back of the department that can only be accessed when the department is open. Researchers can use this computer to work with CD's that are not available on the four PC's in the department's public reference area, for example, lesser used CD ROMs such as the Census CD ROMs for other states or incorporating complex software such as the National Health Interview Survey CD ROMs. This computer is also the only computer with dBase software. Where extractions are too complex to be done easily or efficiently with the vendor produced front end, we provide extraction services using dBase software at this station.

We did not have a formal policy to loan CD ROMs until last year when we instituted a three day loan period for CD ROMs that were not accessed on the public CD ROM stations. We are usually able to loan most of our CDs due to our arrangement with the Law Library. As a second depository on the same campus, the Law Library selected the CD ROMs and transferred them to our department. Two other developments influenced this decision. One was the increase in the number of CD ROMs that we have received in formats not specific to MS DOS environments. These

include flat files, microdata files, and geographic or image files. As we don't have appropriate software in the department to work with these types of data, we allow users to take them to other sites that do. Second was the increase in PCs equipped with CD readers. While it was once rare for end users to have their own CD reader, it is now quite commonplace, and many users would prefer to work with the data on their own systems.

The geographic datafiles represented the largest category of files^{*} for which we did not offer any computer-based access within the department. We have a large number of departments and labs on campus working with geographic files; and while we don't have the facilities to provide GIS services in our department, we are the largest archive of raw geographic data on campus. This includes the Census Bureau's 1990 and 1992 TIGER Line Files, and the U.S. Geological Survey's Digital Line Graph series at 1:2,000,000 and 1:100,000. We are also anticipating the receipt of approximately three hundred fifty CD ROMs representing the U.S. Geological Survey's Digital Orthophotographs for California, geographic image files, in JPEG format.

Our network approach was not part of some grand plan, but rather a large number of circumstances coming together at all at once. We decided in Spring 1994 to ask for a grant for more computing equipment to support access to geographic data files. The Census Bureau was starting to distribute the Landview software with the TIGER Line Files, and we felt that we could not work with this software on an 80386, our most powerful microcomputer. As 80486s were becoming common, we thought that we would ask for this. Our administration increased our request for equipment; it was not coming out of their budget. We ended up with a Pentium, a larger monitor, and a color printer. However, when the equipment arrived, it was still not obvious what software we would use or even what public access we would provide to the equipment.

Before the equipment arrived, the staff support person that I had for computing left to go to another department on campus. We decided that, rather than hire new staff, we would hire a student. When we were able to hire a student who was knowledgeable about both UNIX and networking, we developed our strategy around this person. So we tried Linux on our new equipment. Linux is a freely available UNIX system for Intel based PC's. We decided upon Linux for several reasons. First, because of the complete network support offered; and second, because we had a student capable of installing and maintaining the system. We were also intrigued by the possibility of running GRASS on the system. GRASS is a free GIS system developed by the U.S. Army. GRASS is available for several platforms, including Linux, and is supported by other GIS projects on campus. While we have not worked with GRASS at this point, we are cooperating with GIS labs on campus that use GRASS, and are now examining GRASS as an extraction tool for geographic data.

Our first objective was to provide anonymous FTP access to the system. We started out with two triple speed CD ROM drives, and later replaced them with four double speed CD drives as we decided that the slower CD drives were adequate for throughput on the network. Our most heavily used CD's that could best be accessed in this manner were our 1992 TIGER Line Files for California; these are distributed on three CD's. With four CD drives, we have dedicated three CD ROM drives to TIGER leaving one drive free for other data. Since implementing anonymous FTP access to the TIGER Line Files, we have also made the files available via the World Wide Web using an http front end to the anonymous FTP access. Our CD ROM drives are also available to local users via NFS on an experimental and still restricted basis.

We focused on network access to our large depository data sets for several reasons. First, we were not able to provide adequate access to the data within the department, and several appropriate computing facilities were available on campus. Second, there were competing demands on campus for the TIGER Line Files that could not entirely be met by loaning the CD ROMs: several different researchers would request the same files at the same time, and the researchers frequently did not have adequate access to CD ROM drives in the GIS facilities. Third, network access was readily available in the building. Fourth, appropriate software, i.e. the Linux operating system, was freely available; and finally we were able to hire an experienced student to implement the system.

So far, we have been able to distribute the raw TIGER files to campus users via anonymous FTP. We have been able to make additional CD ROMs available on the system as necessary. We have been able to upload large data extractions created on the 80386 system, the fifth public access microcomputer described above, onto the UNIX system for anonymous FTP access. And finally, we have started to develop an http interface to our anonymous FTP system for access via the WWW.

The networked access to the depository CD ROMs is still very much an experiment, but we have been satisfied with the success of the project so far. We still need to implement a formal system of communication via email to rotate CD ROMs through the fourth CD ROM drive, and once that we conclude that the system is stable we need to increase our efforts to publicize the system.

We do not intend to make a large number of data files available on the network permanently. Our objective in developing this system has been to use the network to serve a local community of data users. However, we have no objections to outside use to the extent that outside use of the system does not compete with local user needs.

* Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

Preserving Scientific Information on the Physical Universe

by Kenneth Thibodeau¹

Background

Over the past fifteen years, there have been several collaborative studies of the archival value of scientific records in the United States. Between 1978 and 1983, representatives of the History of Science Society, the Society of American Archivists, the Society for the History of Technology, and the Association of Records Managers and Administrators worked together on the Joint Committee for the Archives of Science and Technology (JCAST), assessing the state of documentation on research and development, the dissemination of ideas, technology transfer, and professional education in science and technology². A self-acknowledged sequel to the JCAST project, was the collaboration of Joan Haas, Helen Samuels and Barbara Simmons at the Massachusetts Institute of Technology (MIT) which resulted in the publication of *Appraising the records of modern science and technology: a guide* in 1985³. Beginning in 1989, the Center for the History of Physics of the American Institute of Physics (AIP) inaugurated a long term study of fields of physics and related sciences where multi-institutional collaborations are prominent⁴.

While the three projects were undertaken in different organizational contexts, with varying focus and goals, they share a primary concern with the records of research and development activities as resources for historical research. In these endeavors, the potential long-term value of the records of science and technology for further research in these fields themselves has been recognized, but not explored in depth. Generally, consideration of enduring value for science has focused on the data records generated in research and development activities. In fact, JCAST declared, "The first consideration regarding retention of data must be the needs scientists themselves have for these records." Both the JCAST and MIT publications recognized that the actual retention of data for science is usually in the hands of the scientists themselves or of specialized scientific data centers, although archivists may occasionally face the necessity of deciding on the retention of scientific data for scientific purposes⁵. The AIP project took into account the future needs of physicists. It identified categories of records which should be retained by scientific laboratories and science libraries, but did not articulate criteria for identifying records with continuing value for science.

Both the JCAST report and the MIT *Guide* drew attention to the distinction between observational and experimental data, and suggested that long-term value is more often found in

observational data than in experimental results. The argument which supports this generalization is that experiments are repeatable, while observational data of ten relate to unique or rare events or sequences of events. With respect to scientific data, the AIP has concluded that, in high energy physics at least, very little data should be preserved for long periods, and then for purposes of exhibit, rather than scientific research⁶.

Scientific Records in the National Archives of the United States

The National Archives and Records Administration (NARA) of the United States has been involved in appraising and preserving the records of science and technology since its inception. The types of scientific records in the National Archives include project case files, technical reports, laboratory notebooks, drawings and specifications, maps, charts, graphs, aerial photography, motion pictures, sound recordings, and digital data files. The subjects reflect the broad range of scientific and technical activities in which the Government of the United States has engaged, including astronomical, geological and meteorological observations, land and stream classifications, patents, weights and measures, nuclear energy, mineral deposits, weapons systems, aircraft and spacecraft, epidemiology and biometry, entomology, and many other subjects.

It is probably impossible to categorize in general terms the reasons why such scientific and technical records have been accessioned into the National Archives. However, among other factors, NARA has been concerned with the continuing value of these records for science and technology themselves⁷. Concern with the long term scientific value of records derives from a crucial provision of U.S. law, the definition of a federal records. This definition, articulated in title 44 of the United States Code, states that federal records are preserved or appropriate for preservation either as evidence or "because of the informational value of the data in them."

As JCAST recognized, the primary informational value of scientific data is for scientists. In appraising the records of science, then, NARA has an obligation to consider their continuing value to scientists. The most obvious means of exploring this value is to consult with scientists, as was the case in all three of the earlier projects I mentioned. In fact, the reports of all three projects recommended this practice as standard. In the case of research funded by the U.S. Govern-

ment, there are especially important reasons to include the perspective of specialists: (1) the functions of Federal agencies in sponsoring research, (2) the role of researchers outside of the Government in the life-cycle of the data, and (3) the knowledge of the provenance and life-cycle of the data that these scientists have.

(1) Agency Functions: Commonly, Federal records are documents used by agencies in the exercise of mission functions. The Internal Revenue Service, for example, collects tax returns as instruments critical to its function of collecting taxes. Often, however, the science agencies do not use the data that result from the research they sponsor; rather, their function is to sponsor research. In many cases where the funding agency requires the researchers to deliver the resultant data to the Government, the primary, if not sole, purpose for this is to make the data available to yet other researchers outside of the Government, in many cases for long periods of time.

(2) Role of outside researchers: A large proportion of the scientific data generated by the Federal Government is created as a result of the initiatives of investigators outside of the Government. Outside scientists often originate the research proposals and are responsible for the organization and conduct of the research. Through peer review of proposals, other members of the research community have a decisive role in determining what data are collected and how they are collected and organized. Even in cases where research is conducted by scientists employed by an agency, it is not uncommon to have government laboratories reviewed by peer groups composed of outside scientists.

(3) The Life-Cycle of Scientific Data: In many cases, scientific data are in the custody of researchers outside of the government for a major part of their life-cycle. These researchers collect the data, calibrate and refine it, and analyze it, definitively shaping the record. In many cases, the records are not transferred to government custody until the records cease to be active. In other cases, the records - although owned by the government - remain in the custody of outside researchers throughout their life-cycle.

The relevant framework of appraising scientific data sets, thus, is not defined by the business activities or the need for corporate memory of the sponsoring agency, but by the research community. Seeking the input of scientists in the appraisal of the data recognizes that the roles and the actions of academic researchers are at least as important as the functions of the agency that funded the research or launched the satellite.

Since 1990, NARA has sponsored two important efforts to obtain the advice of subject matter experts on the retention of data. The first study, undertaken by the National Academy of Public Administration (NAPA), focused on major federal databases used in support of mission activities. This study

had a twofold purpose: first, to identify these databases and, second, to recommend what data should be preserved in the National Archives⁸.

The NAPA project included the review of some scientific and technical data, notably in the areas of natural resources, the environment, and health. However, large collections of scientific data were intentionally excluded from consideration in the NAPA project because NARA felt that such a large and complex area as "big science" merited separate attention. Scientific data are the focus of the second recent study sponsored by NARA. This project, inaugurated in 1992, was undertaken by the National Academy of Sciences' National Research Council (NRC).

The NRC study was divided into five subject areas: (1) space sciences; (2) physical, chemical and materials sciences; (3) earth sciences, (4) atmospheric sciences, and (5) ocean sciences. Panels of experts were organized to develop recommendations for the preservation of records in each of these areas of research. A steering committee oversaw the work of the panels and formulated generalized recommendations and criteria for the retention of scientific records based on the work of the panels.

The NRC project gave the National Archives the opportunity to interact with the records creators and to engage them in a dialogue on the long term value of the data for secondary use. It was hoped that the NRC project would serve to raise the addressing the complete potential life-cycle of the records during the development and performance of research projects. The final report:

Preserving scientific data on our physical universe. A new strategy for archiving the Nation's scientific information resources, moves towards fulfilling that hope⁹. The report, which was completed in March of this year, makes several sweeping recommendations which can be grouped under the twin headings of retention and responsibilities.

Retention

Recommendation: "As a general rule, all observational data that are nonredundant, useful, and documented well enough for most primary uses should be permanently maintained. Laboratory data sets are candidates for long-term preservation if there is no realistic chance of repeating the experiment, or if the cost and intellectual effort required to collect and validate the data were so great that long-term retention is clearly justified¹⁰."

The report makes two procedural suggestions related to appraisal. The first is that each program or project should have a data management plan established at the origin and governing the entire life-cycle of the data: "Planning activities at the point of data origin must include long-term data management and archiving."¹¹ The second is that appraisal itself is a multifaceted, continuing process:

"Formal appraisals should be kept to minimum, appraisals should be performed according to the data management plan established for each project¹²." The first step in this process would be an interdisciplinary consensus regarding broad classes of data:

"All stakeholders... should be represented in the broad, overarching decisions regarding each class of data¹³."

"Scientists, information technology professionals, data managers, librarians, and archivists must unify their expertise in the establishment of a coherent strategy for end-to-end data and information management¹⁴."

Principal investigators and program managers would then appraise the long term value of individual data sets:

"The appraisal of individual data sets ... should be seen as an ongoing, informal process associated with the active research use of the data, and therefore should be performed by the most knowledgeable about the particular data.... In some cases, they may need to involve an archivist or information resources manager to help with issues of long-term retention.¹⁵"

Finally, the judgements of the primary users would be supplemented with some sort of peer review. The purpose of this review would not be to appraise the data, but to determine if they are as purported and if they are adequately documented. Several options for peer review are identified, ranging from "a formal peer review to certify integrity and completeness," to "documented evidence of the use of the data set in publications in peer-reviewed journals," or evidence from expert users that the data set "is as described in the documentation."

Responsibilities

Recommendation: "As a general principle, data collected by an agency should remain with that agency indefinitely.¹⁶" "Collection" in this context refer to data collected under agency sponsorship, through contracts and grants, as well as data created within the agency. The proposal is for scientific data to be held for the long term in distributed archives, typically in discipline oriented data centers, such as the National Space Science Data Center in NASA and the National Geophysical and Solar-Terrestrial Data Center in NOAA. The novelty in the approach advocated by the report is a proposal for coordination of the activities of such distributed archives "The federal government should create a National Scientific Information Resource Federation — an evolutionary and collaborative network of scientific and technical data centers and archives...."¹⁷

This recommendation to establish the federation suggests action by the Clinton Administration's Information Infrastructure Task Force, the National Science and Technology Council, and/or the Office of Management and Budget to

initiate the federation. The report recommends that either an independent commission or an agency with an established mandate in both the physical sciences and information technology, such as the National Science Foundation, provide executive support for the federation. However, the organization is to be true federation; i.e., a collaboration among equals, not a top-down activity directed by the government.

Hypothetical Profile of the Life Cycle of a Data Set

[In an effort to understand the recommendations of the NRC report, I have constructed the following hypothetical profile of the life-cycle of a data set in accordance with these recommendations. This profile has been reviewed by both the project director and the chair of the steering committee. They both agreed that it accurately reflects the intent of the recommendations.]

When a research project is proposed, the principle investigators, with appropriate collaboration by the program manager in the funding agency, draft a data management plan. In evaluating the long term value of the data, the principals consult with NARA to learn of any recommendations that may have been made by diverse groups of stakeholders about the enduring value of data in the relevant class. The plan is completed no later than project initiation. Following generally accepted criteria, the plan assumes that most observational data produced by the project will be subject to long term retention. Laboratory or engineering data sets, however, are candidates for long-term preservation only if there is no realistic chance of repeating the work, or if the cost and intellectual effort required to collect and validate the data would be so great that long term retention is clearly justified. Other experimental laboratory data or engineering data generally will not need to be retained after completion of the project. The provisions of the plan conform to well established standards for information technology and documentation, as endorsed by the NSIR Federation.

NARA maintains liaison with the sponsoring science agency and, periodically or when necessary, consults with the agency and the investigators to remind them of their responsibilities for long term retention, management, and access.

During the conduct of the research, the investigators and managers will occasionally and informally consider whether the data actually collected merits retention and also whether the history of the project gives rise to any special requirements for documentation.

At the end of the project, the data is deposited in a data center or field archives, as specified in the data management plan. This repository is designated, or operated, by the lead agency in the subject area, where staff have appropriate expertise and close ties with the relevant researcher community. The repository has mechanisms for access to the data by individuals beyond the primary users.

For the data set to be accepted into the data center of distributed archives, it must undergo some form of peer review to ensure that it adequately meets the standards of uniqueness and accessibility. Along with the data, metadata essential for others to use the data is transferred to the repository. When any required services related to retention or access to the data are available elsewhere in the Federation at lower costs than would be incurred by the organization with primary responsibility, those alternative services are used when feasible.

From the time the data set becomes available for use outside of the project, it is identified in a hierarchical information locator system. The data should be available for remote access and/or file transfer, ideally as an extension of the locator system.

NARA is informed of the existence and location of the data. NARA monitors the preservation and accessibility of the data over time and, acting in an advisory capacity, helps the custodians with any problems in these areas. If the custodians can no longer meet the needs of the user community, or if the data is no longer in regular use, the data should be considered for transfer to some other federal science agency or, as a last resort, to the National Archives.

The NRC report has been received too recently for NARA to have been able to take a position concerning its recommendations. However, as an individual archivist, I would like to offer some observations. They are entirely my own; they do not represent the position of the National Archives not, as far as I know, of any other individual in the National Archives.

Preserving scientific data on our physical universe purports to offer "a new strategy for archiving the Nation's scientific information resources." There are certainly new elements in this strategy. One is the articulation of the general principle that data collected in the observational sciences should be preserved permanently. Another is the creation of the NSIR Federation, conceived as a collaboration facilitated, but not directed, by the Federal Government. A major innovation entailed by the recommendations is that the scientific community would have to recognize data management, data retention and data access as valuable activities by scientists, and would have to adjust the culture of science to include rewards for these activities, on a par with the publication of research results.

These things would be significant changes, but they would be changes in the scientific community. From an archival perspective, the impact would be different. As the report recognizes, very little scientific data has been deposited in the National Archives, and there is no grounds for expecting this to change. The assertion that entities within the scientific community should be responsible for the long-term retention of scientific data sets and for access to them is not only consonant with the status quo, but also consistent with

archival concepts, as articulated by the three projects I described at the start of this paper. Furthermore, the National Archives does not play a forceful role in the management, retention, or access to scientific data. The report, in fact, argues that such a role would be counter-productive. It suggests that the appropriate role for NARA would be that of consultant and collaborator on archival preservation and access issues. On empirical grounds, one might say that this is role that NARA has been playing in the domain of scientific data. Thus, one might argue that, from an archival perspective, the recommendations of the NRC report reduce, by and large, to a confirmation of the status quo.

References

- 1 The views stated in this paper are those of the author and do not represent the position of the National Archives and Records administration. (Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.)
- 2 Clark A. Elliott, editor, Understanding progress as process. Documentation of the history of post-war science and technology in the United States. Final Report of the Joint Committee for the Archives of Science and Technology. Chicago. Society of American Archivists, 1983.
- 3 Joan K. Haas, Helen Willa Samuels and Barbara Trippel Simmons, Appraising the records of modern science and technology: a guide. Cambridge, MA. Massachusetts Institute of Technology, 1985.
- 4 Joan Warnow-Blewett and Spencer Weart, AIP Study of Multi-Institutional Collaborations, Phase I: High-Energy Physics. Report No. 1: Summary of Project Activities and Findings, Project Recommendations. New York: American Institute of Physics, 1992.
- 5 Elliott, p. 33-34. Haas et al., pp. 60-61.
- 6 Joan Warnow-Blewett, Lynn Maloney, and Roxanne Nilan, AIP Study of Multi-Institutional Collaborations, Phase I: High-Energy Physics. Report No. 2: Documenting Collaborations in High-Energy Physics. New York: American Institute of Physics, 1992. Pp. 75-76, 89-91.
- 7 Trudy Huskamp Peterson. Presentation on the National Archives and the Records of Science for the National Academy of Science/National Research Council Study on the Long Term Retention of Scientific and Technical Records. Plenary Session, July 7, 1993.
- 8 National Academy of Public Administration, The Archives of the Future: Archival Strategies for the Treatment of Electronic Databases. A report for the National Archives and Records Administration. Washington, D.C. National Academy of Public Administration, 1991.
- 9 National Research Council, Preserving scientific data

on our physical universe. A new strategy for archiving the Nation's scientific information resources. Commission on Physical Sciences, Mathematics, and Applications. Washington D.C. National Academy Press. 1995.

10 Ibid. p. 4. The report argues that technological developments make it possible to save everything and to provide access to it. However, the report recognizes that data management activities in general, not just preservation, are chronically underfunded and that they are at best a secondary concern in scientific culture. it does not adequately address how these difficulties can be overcome.

11 Ibid. p. 50

12 Ibid. p. 40

13 Ibid.

14 Ibid. p. 50

15 Ibid. p. 4

16 Ibid. p. 56

17 Ibid. p. 51



INTERNATIONAL ASSOCIATION FOR
SOCIAL SCIENCE INFORMATION
SERVICE AND TECHNOLOGY

• • • • •
ASSOCIATION INTERNATIONALE
POUR LES SERVICES ET
TECHNIQUES D'INFORMATION EN
SCIENCES SOCIALES

Membership form

The **International Association for Social Science Information Services and Technology (IASSIST)** is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from re-

duced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:

Regular Membership: \$40.00 per calendar year.

Student Membership: \$20.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription:

\$70.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of
IASSIST. Please see my choice below:

- ☐ \$40 Regular Membership
☐ \$20 Student Membership
☐ \$70 Institutional Membership

My **primary** Interests are:

- ☐ Archive Services/Administration
☐ Data Processing
☐ Data Management
☐ Research Applications
☐ Other (specify) _____

Please make checks payable
to **IASSIST** and Mail to :
Mr. Marty Pawlocki
Treasurer, **IASSIST**
% 303 GSLIS Building,
Social Science Data
Archives, University of
California, 405 Hilgard
Avenue, Los Angeles, CA
90024-1484

Name / title _____

Institutional Affiliation _____

Mailing Address _____

City _____

Country / zip/ postal code / phone _____

REQUESTED

JUL 18 '97



U.S. GOVERNMENT

1.01

OH

6819820

SERIALS DEPARTMENT
UNIV OF NORTH CAROLINA-CHAPEL HILL
CB #3938 DAVIS LIBRARY
CHAPEL HILL, NC 27514-8396

555 UC XL 179
12/29/97 32590